

University of Groningen

## Data-driven identification of fixed expressions and their modifiability

Villada Moirón, María Begoña

**IMPORTANT NOTE:** You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

*Document Version*

Publisher's PDF, also known as Version of record

*Publication date:*

2005

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Villada Moirón, M. B. (2005). Data-driven identification of fixed expressions and their modifiability s.n.

**Copyright**

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

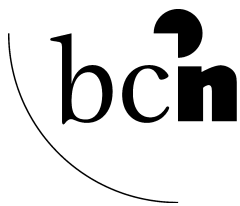
**Take-down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Data-driven identification of fixed expressions  
and their modifiability

María Begoña Villada Moirón



The work in this thesis has been carried out under the auspices of the Behavioral and Cognitive Neurosciences (BCN) research school, Groningen, and has been part of the PIONIER project *Algorithms for Linguistic Processing* supported by grant number 220-70-001 from the Netherlands Organization for Scientific Research (NWO).



Groningen Dissertations in Linguistics 52

ISSN 0928-0030

Document prepared with L<sup>A</sup>T<sub>E</sub>X 2<sub>ε</sub>

Cover designed by Misha Scholten. Printed by PrintPartners Ipskamp.

Rijksuniversiteit Groningen

Data-driven identification of fixed expressions and their  
modifiability

Proefschrift

ter verkrijging van het doctoraat in de  
Letteren  
aan de Rijksuniversiteit Groningen  
op gezag van de  
Rector Magnificus, dr. F. Zwarts,  
in het openbaar te verdedigen op  
donderdag 24 maart 2005  
om 13.15 uur

door

María Begoña Villada Moirón

geboren op 11 juli 1972  
te Riotorto, Spain

Promotor: Prof. dr. ir. J. Nerbonne

Copromotores: Dr. G. van Noord  
Dr. G. Bouma

Beoordelingscommissie: Prof. J. Hoeksema  
Prof. J. Odijk  
Prof. I. Sag

# Preface

During the years that I spent as a PhD student at the Alfa-Informatica department I met many people to whom I am truly grateful and who had a major impact on my thesis, my life, or both.

This thesis would never have come into being without the invaluable advice, guidance and support of my two supervisors: Gosse Bouma and Gertjan van Noord. They always found time to listen to the many difficulties I encountered along the way and discuss solutions. From countless discussions with them, I learnt a lot about scientific research, statistics, computational linguistics and (aghhh, I even have to mention them in the preface!) fixed expressions. Thank you both for your crucial input and the many corrections of former drafts. And, thank you for letting me work at my own pace and pushing me when necessary.

I owe a heavy debt to my promotor John Nerbonne whose wisdom and suggestions jotted down in numerous drafts I tried to adopt, although not always succeeding in covering them all. Throughout the whole project, John helped me with the statistics; he also read drafts amazingly quickly and returned insightful feedback, as well as corrected my English writing. Of course, thank you for providing extra financial support when needed and for maintaining a healthy research environment.

To the list of colleagues deserving of gratitude, I'd like to add Leonoor van der Beek who always agreed to read my writings and returned criticism and interesting suggestions for improvement. Leonoor always found a place in her agenda to discuss my work and the so much needed linguistic data. I also owe her for the Dutch translation of the thesis summary. Thanks also to Lonneke van der Plas for her contribution.

Being a non-native and clumsy Dutch speaker posed difficulties to evaluate the output of my experiments. Without hesitation, Gosse Bouma, Gertjan van Noord, and Leonoor van der Beek, stoically assessed long lists of (sometimes nonsensical) expressions and always gave me constructive feedback. I'm also grateful to Robbert Prins, Henny Klein, Jack Hoeksema, Bart Hollebrandse and Dirk-Bart den Ouden for assessing and discussing the

specific language data.

I am also indebted to the members of my reading committee: Jack Hoeksema, Jan Odijk and Ivan Sag. They corrected various mistakes in a previous manuscript and made detailed and helpful suggestions for improvement that (to some extent) were incorporated in this book. None of the above is of course responsible for how the thesis turned out in the end; the remaining mistakes or inaccuracies are entirely my own responsibility.

Many other people also helped me in many ways. I would like to acknowledge the members of the PIONIER project *Algorithms for Linguistic Processing* for all their instruction and discussions, in particular its former members Tanja Gaustad, Jan Daciuk and Rob Malouf. My work also benefited from the input of many colleagues who participated in various reading groups and contributed to many fruitful exchanges. In addition, I am grateful to my (through-various-years) office mates: Rob Malouf, Stasinou Konstantopoulos, Susanne Schoof, Holger Hopp and Jörg Tiedemann for their assistance with a variety of technical matters, thought-provoking discussions, or simply, their occasional interruptions to remind me to take a coffee break. Stas deserves separate mention for creating the RuG thesis style file. A big thanks to Mark-Jan Nederhoff for sharing his substantial knowledge of L<sup>A</sup>T<sub>E</sub>X and book typesetting. I also would like to acknowledge the encouragement and support of my colleagues in Alfa-Informatica who make the department such a stimulating and enjoyable environment for research. Last but not least, I thank the fourth floor secretarial staff for their assistance with the day-to-day issues and our system administrators for their computer support.

Among those less directly related to this work, I am grateful to Carl Vogel who encouraged me to continue doing research in computational linguistics, even if that meant abandoning the at-the-time ‘dreadful’ *clitics*. Also, thanks to the Hungarian members of the OTKA-NWO project *Finding and Processing Multi-word Lexemes* for testing my methods on a totally different language, with different settings and application, as well as for many hours involved in e-discussions.

Extra special thanks to Leonoor van der Beek and Francisco Borges who kindly agreed to be my ‘paranimfs’.

Como sempre, agradezo inmensamente a axuda da miña familia e é unha honra que poidan asistir á miña defensa de doutoramento. Finally I owe my greatest debt to Michiel for his help, support and willingness to be distracted from his own studies and work by talk of statistics and linguistics.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.1.1	Chapter overview . . . . .	1
1.1.2	What are <i>fixed expressions</i> ? . . . . .	2
1.1.3	Regularity, semi-regularity and exceptions in the grammar and lexicon . . . . .	3
1.2	Research questions . . . . .	8
1.2.1	Aims and objectives . . . . .	9
1.2.2	Practical motivation . . . . .	10
1.3	Linguistics, Corpora and Statistics . . . . .	10
1.3.1	Previous linguistic descriptions . . . . .	11
1.3.2	Corpus-based approaches . . . . .	12
1.4	This thesis . . . . .	15
1.4.1	Characteristics of fixed expressions . . . . .	15
1.4.2	Identification of fixed expressions . . . . .	17
1.4.3	Exploring variation and modifiability . . . . .	18
1.5	Applications . . . . .	19
1.6	Chapter summaries . . . . .	21
<b>2</b>	<b>On fixed expressions</b>	<b>23</b>
2.1	Introduction . . . . .	23
2.2	Semantic properties of fixed expressions . . . . .	24
2.2.1	Denoting and non-denoting words . . . . .	25
2.2.2	Opacity and conventionality . . . . .	26
2.2.3	Decomposability . . . . .	31
2.3	Lexical and morphological properties . . . . .	34
2.3.1	Selectional restrictions and lexical fixedness . . . . .	34
2.3.2	Idiom families . . . . .	35
2.3.3	Restricted morphological operations . . . . .	36
2.4	Syntactic properties . . . . .	38
2.4.1	Sense syntax asymmetry . . . . .	43



2.4.2	Syntactic versatility . . . . .	44
2.5	Summary . . . . .	47
<b>3</b>	<b>Automatic extraction methods</b>	<b>49</b>
3.1	Introduction . . . . .	49
3.1.1	Overview of identification models . . . . .	49
3.1.2	Purely statistical approach . . . . .	50
3.1.3	Hybrid approaches: linguistics and statistics . . . . .	52
3.2	Pre-processing extraction data . . . . .	53
3.2.1	Extraction data . . . . .	53
3.2.2	Annotation tools for Dutch . . . . .	54
3.2.3	Sampling and representation . . . . .	54
3.3	Collocation statistics . . . . .	56
3.3.1	Contingency table . . . . .	56
3.3.2	Association measures . . . . .	57
3.3.3	Hypothesis testing . . . . .	61
3.4	Loglinear models . . . . .	62
3.5	Evaluation methodology . . . . .	63
3.5.1	Validation data . . . . .	64
3.5.2	Quantitative evaluation . . . . .	65
3.6	Choosing the best approach . . . . .	65
<b>4</b>	<b>Identification of Collocational Prepositional Phrases</b>	<b>69</b>
4.1	Introduction . . . . .	69
4.2	Linguistic properties . . . . .	71
4.2.1	Idiosyncratic features . . . . .	72
4.2.2	Discussion . . . . .	76
4.3	Extracting CPPs from a corpus . . . . .	76
4.3.1	Resources . . . . .	76
4.3.2	Dataset extraction . . . . .	77
4.4	Collocation statistical tests . . . . .	78
4.4.1	Bigram model . . . . .	78
4.4.2	Trigram model . . . . .	80
4.5	Evaluation and results . . . . .	81
4.5.1	Methodology . . . . .	81
4.5.2	Bigrams results . . . . .	83
4.5.3	Trigrams results . . . . .	84
4.6	Discussion . . . . .	84
4.6.1	Identifying the best association measure . . . . .	85
4.6.2	Error analysis . . . . .	87
4.6.3	Effects of having more validation data . . . . .	89

4.7	Summary . . . . .	90
<b>5</b>	<b>Identification of Support Verb Constructions</b>	<b>91</b>
5.1	Introduction . . . . .	91
5.1.1	Support verb constructions . . . . .	92
5.1.2	Aims and overview . . . . .	94
5.2	Candidate extraction from corpora . . . . .	95
5.2.1	Preprocessing . . . . .	95
5.2.2	Building datasets . . . . .	97
5.3	Identification process and evaluation . . . . .	101
5.3.1	Bigram model . . . . .	101
5.3.2	Imposing a frequency cutoff . . . . .	102
5.3.3	Evaluation methodology . . . . .	102
5.4	Quantitative results . . . . .	105
5.4.1	Bare models . . . . .	106
5.4.2	Adding a frequency cutoff: pros and cons . . . . .	107
5.4.3	A few remarks on the loglinear model's performance . . . . .	114
5.4.4	Experiment's validation: other support verbs . . . . .	114
5.4.5	Discussion . . . . .	117
5.5	Qualitative results . . . . .	119
5.5.1	Qualitative divergence between best tests . . . . .	120
5.5.2	Error analysis . . . . .	120
5.5.3	Updating gold-standards . . . . .	122
5.5.4	Discussion . . . . .	122
5.6	Conclusions . . . . .	122
<b>6</b>	<b>Additional Linguistic Diagnostics</b>	<b>125</b>
6.1	Introduction . . . . .	125
6.1.1	Sources of noise in nbest list . . . . .	126
6.1.2	Overview . . . . .	127
6.2	Diagnostics and evidence . . . . .	127
6.2.1	NP pronominalization . . . . .	127
6.2.2	Scrambling . . . . .	128
6.2.3	PP over verb and nominalization . . . . .	129
6.2.4	Coordination . . . . .	131
6.2.5	Some additional remarks . . . . .	132
6.3	Applying diagnostics semi-automatically . . . . .	132
6.3.1	Data resources and tools . . . . .	132
6.3.2	Method . . . . .	135
6.3.3	Preliminary results . . . . .	137
6.4	Evaluation . . . . .	138

6.4.1	Methodology . . . . .	138
6.4.2	Results . . . . .	138
6.5	Discussion . . . . .	139
6.5.1	Efficacy of the linguistic diagnostics . . . . .	139
6.5.2	Annotated data, search queries and evidence . . . . .	141
6.6	Conclusions . . . . .	145
<b>7</b>	<b>Variation within Support Verb Constructions</b>	<b>147</b>
7.1	Introduction . . . . .	147
7.1.1	Mining linguistic descriptions from corpora . . . . .	149
7.1.2	Overview . . . . .	150
7.2	Modification and variation types . . . . .	150
7.2.1	Evidence we seek . . . . .	151
7.3	A corpus-based method . . . . .	155
7.3.1	Settings . . . . .	155
7.3.2	The extraction process . . . . .	157
7.3.3	The evidence retrieved . . . . .	159
7.3.4	Evaluation . . . . .	161
7.3.5	Summary . . . . .	163
7.4	Assessing modifiability . . . . .	164
7.4.1	Trends in extracted evidence . . . . .	166
7.4.2	Informative determiner changes . . . . .	170
7.4.3	Underlying semantic structure . . . . .	171
7.5	Summary . . . . .	173
<b>8</b>	<b>Conclusions and Future directions</b>	<b>175</b>
8.1	Future directions . . . . .	178
	<b>Bibliography</b>	<b>181</b>
	<b>A Validation data</b>	<b>191</b>
	<b>B Dutch fixed expressions in text</b>	<b>195</b>
	<b>Samenvatting</b>	<b>199</b>
	<b>Groningen Dissertations in Linguistics</b>	<b>205</b>